(12) UK Patent Application (19) GB (11) 2 264 186 (13) A

(43) Date of A publication 18.08.1993

(21) Application No 9202288.8

(22) Date of filing 04.02.1992

(51) INT CL[5]
G06F 15/40

(52) UK CL (Edition L)
G4A AUDB
G4H HTAT H13D H14A
U1S S2268

(56) Documents cited
EP 0241646 A2 US 4774661 A

(58) Field of search
UK CL (Edition K) G4A AUD, G4H HTAE HTAT
INT CL[5] G06F 15/40
On-line database: WPI

(71) Applicant
Inishashe Limited

(Incorporated in Ireland)

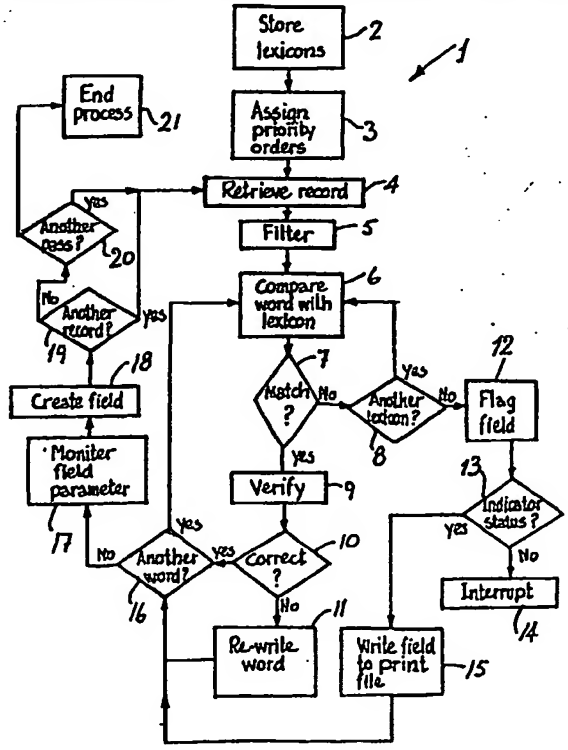
41 Fitzwilliam Street, Dublin 2, Ireland

(72) Inventor
Martin Biddle

(74) Agent and/or Address for Service
Marks & Clerk
57–60 Lincoln's Inn Fields, London, WC2A 3LS, United Kingdom

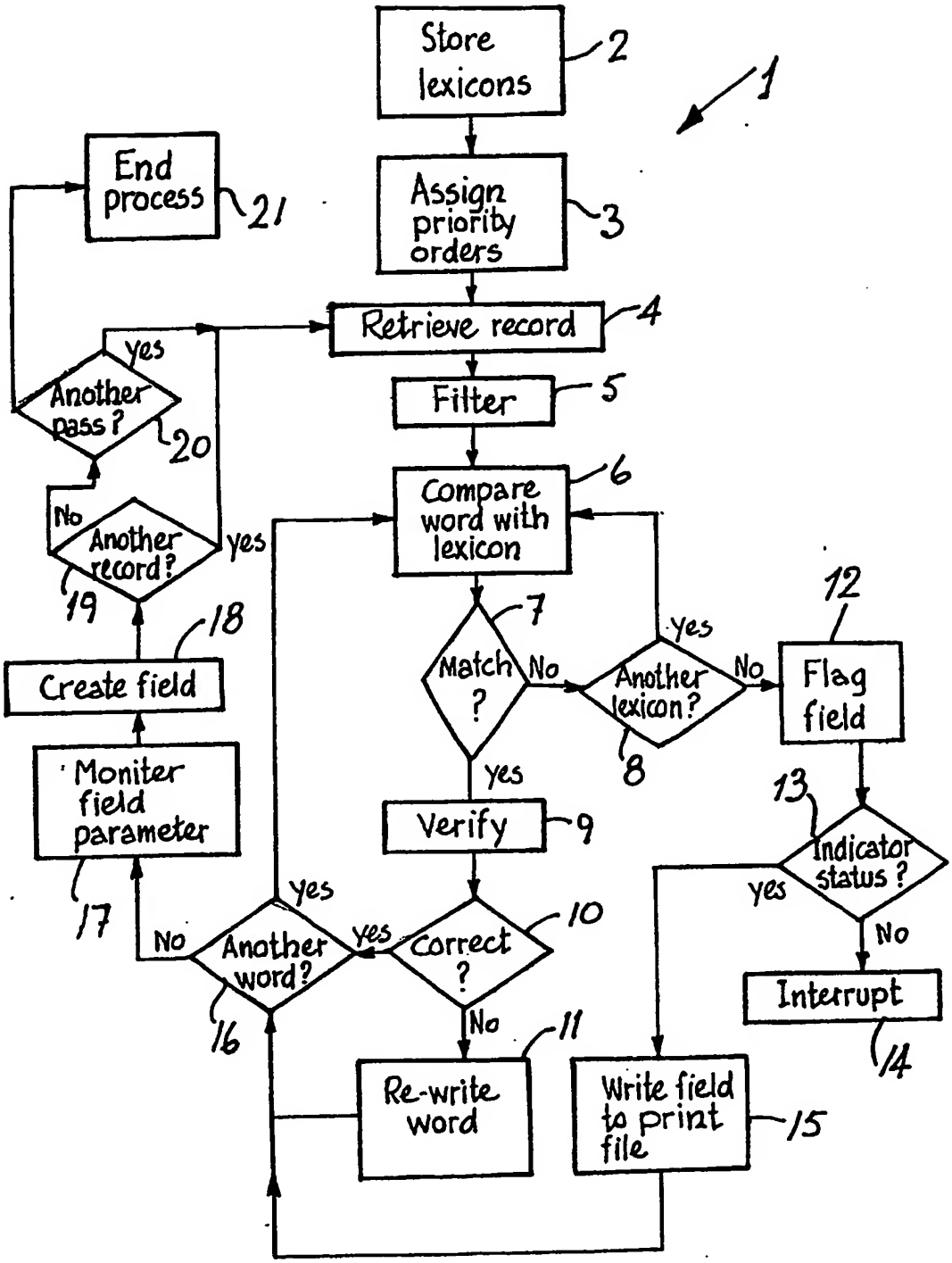
(54) Database correction and conversion

(57) A process 1 is disclosed for conversion of an input database to an output database having an improved data structure and corrected data words. A set of lexicons is stored, 2, for each field of the input database. For each word position in a particular field, a search priority order is set, 3, for the associated lexicons based on the probability of the lexicon containing a word at that position. Each lexicon contains reference words and associated control rules. When a processor matches a word in a field with a lexicon, 7, it verifies the word, 9, by use of the control rules. If incorrect, the word is re-written, 11. If no match is found in the lexicon set a flag is inserted in the field, 12. According to an indicator, 13, processing may be interrupted, 14, for manual input of an instruction, or the flagged field may be written to a print file, 15, for later attention. Efficiency is considerably improved by use of a number of lexicons as opposed to a full data dictionary, by assigning lexicon priority orders, and by insertion of additional fields, 18, according to the contents of existing fields, 17.



At least one drawing originally filed was informal and the print reproduced here is taken from a later filed formal copy.

GB 2 264 186 A


Store lexicons
2
1
End process
21
Assign priority orders
3
Retrieve record
4
Filter
5
Another pass?
20
yes
Compare word with lexicon
6
No
Another record?
yes
19
Match ?
7
No
Another lexicon?
yes
No
12
Flag field
18
Create field
8
Moniter field parameter
Verify
9
13
Indicator status ?
yes
No
17
No
Another word?
yes
yes
Correct ?
10
Interrupt
16
No
11
14
Re-write word
Write field to print file
15

"A Database Conversion Process"

The invention relates to a process for the conversion of an input database which has records with fields of alphanumeric data to an output database which facilitates operation of a processor in retrieval of data and direction of document printing using the data in an error-free and efficient manner.

Such a conversion process would be required, for example, when an input database of names and addresses is to be used for printing of letters in a mail-shot. Names and addresses within the input database may be spelt incorrectly, may be in the wrong position, or may be missing.

Heretofore, the approach to conversion of such an input database has been to use a spell check apparatus such as that described in British Patent Specification No. 2,201,274 B (Brother) which checks spelling of each word in sequence. For a database conversion process, this approach is extremely time-consuming as each individual word must be checked against a large stored dictionary. Further, inaccuracies due to wrong

positioning of words within fields of the input database are not corrected.

The invention is directed towards providing a database conversion process which operates efficiently and which converts an input database so that the data content and structure is consistent and correct and may be used for the printing of documents.

According to the invention, there is provided a process for conversion of an input database having records with fields of alphanumeric data to an output database having a data content and a structure which facilitates operation of a processor in retrieval of data and direction of document printing using the data in an error-free and efficient manner, the process comprising the steps of:-

writing a set of lexicons to a memory circuit, the set of lexicons being associated with a particular field of the records of the input database, and each lexicon having a set of reference data words and associated rules;

assigning a priority order for the set of lexicons for each word position, the priority order being based on the probability of the lexicon containing a reference word for that position of the field;

sequentially writing each record of the input database into the memory circuit and filtering out the particular field; and

for each data word of each field, carrying out the steps of:-

comparing the word with at least some of the lexicons in the priority order until a match is found;

when a match is found, verifying the word by comparing parameters of the word with rules of the reference word in the lexicon, and re-writing the word in the field according to the rules; and

if no match is found inserting a flag in the field.

In one embodiment, the process comprises the further step of of pre-setting a processing indicator, and according to status of the indicator, when a flag is inserted in a field, either interrupting processing and prompting manual user input of an instruction, or writing the field to a printing file for printing of an error list at a later time.

The invention will be more clearly understood from the following description of some preferred embodiments thereof, given by way of example only with reference to the accompanying drawing which is a flowchart illustrating a database conversion process of the invention.

Referring to the drawing, there is illustrated a database conversion process of the invention, indicated generally by the reference numeral 1. The process 1 may be carried out on any suitable computerised database system having a processor, a storage device, a memory circuit, a display unit and a printer. Irrespective of the type of computerised database system used, the technical features of the process lead, as is explained below, to improved efficiency in database conversion and to significantly fewer errors in the output database. Indeed, it is quite possible to eliminate errors completely using the process 1.

The first step of the process is step 2, which involves storing in the storage device a number of lexicons. Each lexicon is a set of reference words, such as in a dictionary, an in addition a set of rules associated with each reference word. There is one lexicon for each "type" of data which is likely to be found in an input database. For example, one lexicon may relate to salutations for names in a database. In such a lexicon, the word "Esq.," will be stored as a reference

word and a rule indicating that it should be removed from the database and the salutation "Mr." should be inserted as a first word in the name field instead. In general, not all possible surnames or Christian names will be included in the one lexicon. It is more efficient to create many different lexicons, each having a portion of the total number of possible names, sorted in alphabetical order. This is possible because when the input database is being processed, it would be in alphabetical order of the names in the data records. Other examples of lexicons include a postal district number lexicon and a street name lexicon. In these examples, rules in the postal district lexicon will indicate a number of possible street names for that postal district. Creation and storing of the lexicons in step 2 is an extremely important part of the process and it must be carried out with the input database in mind. In general, there will be a set of lexicons created for each type of input data.

Step 3 of the process is also quite important as it involves taking the set of lexicons which has been created for any particular input database and assigning a priority order to it for each data position within fields of the input database records. For example, the priority order for the first word in a name field will indicate that the salutation lexicon is of highest priority and the relevant surname lexicon as the second highest priority. Each word "position" within the input database fields has a priority order. For example, the

priority order for the third position in a name field would assign highest priority to a Christian name lexicon. Thus, for each set of lexicons (associated with an input database) there are a number of different priority orders, each priority order being associated with a word or data position within a field of the input database records.

In step 4, the input database stored on the storage device is accessed and a record is retrieved and loaded into the memory circuit. In step 5, the processor filters out the relevant field which this particular path of the process is concerned with. This field may, for example, be the name field in each record. In step 6 the processor reads the first word of the field and compares it with the lexicon which has been assigned as the highest priority order for the first word position in the field. In this example, the salutation lexicon is the first chosen. In step 7, the process determines whether or not there is a match with the lexicon, i.e. whether or not there is a reference word the same as the first word which has been read from the field. If not, in step 8 the processor determines if there is another lexicon in the priority order and in step 6 repeats the process of comparing the word with the lexicon of next priority order. An example of where there will be no match for the first lexicon is if the salutation "Esq." is included at the end of the name field and there is no salutation at the beginning. In this case, the first word to be read will be a surname and this is likely to be found in

the second lexicon. However, in most cases the salutation "Mr.", "Ms." or "Mrs." will be used and that is why the salutation lexicon was assigned highest priority. If a match is found in a lexicon, i.e. if the word which has been read is the same as the reference word in a lexicon, in step 9 the processor verifies the word by monitoring the control rules associated with the reference word. These control rules may indicate that the words should be in a particular position in the name field. For example, if the salutation "Esq." was included at the end of a name field in the input database, the first word would be the surname and the control rule will indicate that it should be in the second word position of the name field instead of the first. In step 10 the processor determines according to the control rules if the word in the input database is correct and in step 11 the word is re-written in the field, for example, the surname is re-written into the second word position in the field.

Returning now to step 8 which involves the processor determining if there is another lexicon in the priority order for that word position, if all lexicons in a priority order have been accessed and no match has been found, in step 12 a flag is inserted in the field. When this has been done, the processor checks the status of a processing indicator which has been set before the process began. This indicator has only two levels, indicated as "1" and "0". In this example, if the indicator is "1", processing is interrupted and the

processor generates a prompt for display on the video screen indicating to a user that no match has been found and the manual instructions at the keyboard are required in order to proceed. On the other hand, if the indicator status is "0", the processor writes the field to a print file and then continues with processing for the next word in the field. At a later time, the print file is used for printing of a list of fields to be manually edited.

Step 16 of checking whether or not there is another word in the field is carried out after either steps 10 (if the previous word is correct), after step 11 (if incorrect) or after step 15. If there is another word in the field which has been loaded into the memory circuit, steps 6 to 16 are repeated for that word. If all words in a field have been verified, in step 17 the processor monitors certain parameters of the field. In one example, the salutation is the parameter which is monitored as an indication as to whether or not the person whose name in the field is male or female. In step 18 another field is created in the record and in this example, the additional field would be a male/female indicator. Thus, not only does the process involve correction of a database and changing the structure to a correct format, but it also involves addition to the database so that it may be processed more easily at a later date, possibly for other uses.

In step 19, the processor checks if there is another record in the input database, and if so steps 4 to 19 are repeated for the next record.

When all records have been processed, in step 20 the processor checks user instructions as to whether or not another pass is required, i.e. whether or not another field within the records of the input database must be processed. For example, one pass would relate to correction of the name field in an input database, whereas another pass would correct the address field. If there is another pass, steps 4 to 20 are repeated for each record. If another pass is not required, the process ends in step 21.

It will be appreciated that by the manner in which sets of lexicons are created, in which priority orders are assigned to them and by which the processor operates in verifying pieces of data using the selected lexicons, processor efficiency is considerably improved for conversion of an input database. Further, accuracy in the output database is considerably improved because the processor may react in an intelligent manner for each different piece of data using the in-built control rules associated with each reference word in the lexicons. Further, reference words which would occur rarely are not included in the lexicons and this saves time for each pass of that particular lexicon. When such a word does occur in an input database processing may be interrupted in step 14

and a manual instruction received. It has been found that this is much more efficient than trying to cater for every possible eventuality automatically. It has also been found that by monitoring field parameters and inserting an additional field in the record, the output database is considerably improved and may be easily used for further processing work which would involve separating records into different sets according to the selected parameter. The selected parameter could be age, postal district, sex or any other parameter.

It has been found to be convenient to set the processing indicator status to "1" for day-time processing where a user is available for inputting of manual instructions, and to "0" for night-time processing, in which case a user would input instructions from a document which has been printed from the print file. This lends considerable versatility to the process of the invention.

The invention is not limited to the embodiments hereinbefore described, but may be varied in construction and detail.

CLAIMS

1. A process for conversion of an input database having records with fields of alphanumeric data to an output database having a data content and a structure which facilitates operation of a processor in retrieval of data and direction of document printing using the data in an error-free and efficient manner, the process comprising the steps of:-

writing a set of lexicons to a memory circuit, the set of lexicons being associated with a particular field of the records of the input database, and each lexicon having a set of reference data words and associated rules;

assigning a priority order for the set of lexicons for each word position, the priority order being based on the probability of the lexicon containing a reference word for that position of the field;

sequentially writing each record of the input database into the memory circuit and filtering out the particular field; and

for each data word of each field, carrying out the steps of:-

comparing the word with at least some of the lexicons in the priority order until a match is found;

when a match is found, verifying the word by comparing parameters of the word with rules of the reference word in the lexicon, and re-writing the word in the field according to the rules; and

if no match is found inserting a flag in the field.

2. A process as claimed in claim 1, comprising the further step of pre-setting a processing indicator, and according to status of the indicator, when a flag is inserted in a field, either interrupting processing and prompting manual user input of an instruction, or writing the field to a printing file for printing of an error list at a later time.

3. A process as claimed in claims 1 or 2, comprising the further steps of writing an additional field to a record in response to monitoring a parameter of a field in the record, the parameter being monitored by reference to a

reference word and control rule associated with a word and a field.

4. A process substantially as hereinbefore described with reference to and as illustrated in the accompanying drawings.

Patents Act 1977
Examiner's report to the Comptroller under Section 17 (The Search Report)

Application number
9202288.8

Relevant Technical fields

(i) UK Cl (Edition K) G4A (AUD); G4H (HTAE, HTAT)

(ii) Int CL (Edition 5) G06F 15/40

Search Examiner
B G WESTERN

Databases (see over)

(i) UK Patent Office

(ii) ON-LINE DATABASE: WPI

Date of Search
14 APRIL 1992

Documents considered relevant following a search in respect of claims 1-4

Category (see over)	Identity of document and relevant passages	Relevant to claim(s)
A	EP 0241646 A2 (TOSHIBA) Whole document	1-4
A	US 4774661 A (KUMPATI) Whole document	1-4

kg - c:\wp51\doc99\fil000953

Category	Identity of document and relevant passages	Relevant to claim(s)

Categories of documents

X: Document indicating lack of novelty or of inventive step.

Y: Document indicating lack of inventive step if combined with one or more other documents of the same category.

A: Document indicating technological background and/or state of the art.

P: Document published on or after the declared priority date but before the filing date of the present application.

E: Patent document published on or after, but with priority date earlier than, the filing date of the present application.

&: Member of the same patent family, corresponding document.

Databases: The UK Patent Office database comprises classified collections of GB, EP, WO and US patent specifications as outlined periodically in the Official Journal (Patents). The on-line databases considered for search are also listed periodically in the Official Journal (Patents).

Published 1993 at The Patent Office, Concept House, Cardiff Road, Newport, Gwent NP9 1RH. Further copies may be obtained from

THIS PAGE BLANK (USPTO)